**学士学位本科生**

**开题报告**

题目：基于动态视觉传感器的手势识别系统

**学 号： 1609030225**

**学科专业： 软件工程**

**年 级： 2016级**

**姓 名： 王建**

**指导教师： 郭磊**

**报告时间：**

**中 国 石 油 大 学**

# 开题报告

# 1．选题目的及研究意义

## 1.1 选题目的

神经形态视觉使用硅视网膜传感器采集数据，例如动态视觉传感器（DVS相机）。动态视觉传感器受到生物视觉的启发，产生的异步事件流指示摄像头视场内各个点的亮度变化。因此，他们能够快速的捕捉变化，天然的消除冗余的背景数据，极大地减少了要处理的数据量，并且它们的动态特性使其非常适合诸如光流，对象跟踪，动作识别或动态场景理解之类的领域。相比传统相机，DVS相机的具有帧速高的特点，能够达到20k FPS的量级，更能够捕捉高速运动和复杂光照条件下的物体。通过动态传感器（DVS相机）可以实时捕捉突然出现的物体目标，同时能够有效地降低延迟，使得手势识别能够在更多的领域（家电、手机等市场）得以应用。

手势是人类交流和表达的一个自然而直观的部分。我们几乎不需要额外的智能处理就能够使用手势与周边环境中的设备进行通信。在最近几年，深度学习得到了快速发展，并且在机器视觉，自然语言理解等上取得了很好的成效。现有的 AlexNet、GoogLeNet 等算法已经在图像分类上取得了很好的成果，这说明卷积神经网络在机器视觉领域具有良好的应用前景，已成为许多科学研究领域的研究热点之一。

## 1.2 研究意义

本课题面向的动态视觉传感器在机器视觉领域十分重要。卷积神经网络在图像处理方面的突破近年来吸引了大量研究者关注这一领域。但是对于高速运动的物体，使用传统相机需要以几千或几万帧/秒的速率才能捕捉到，相机处理能力与计算能力的不平衡的瓶颈，限制了其进一步向纵深发展。因此在图像处理方面的优化从而加速机器视觉发展的工作亟待解决。同时，类似的方法同样也可稍微改进以应用于深度学习的其他方面。

动态视觉传感器的速度不受传统的曝光时间和帧速率限制，像素的相应时间在纳秒级；更重要的是，这种传感器可以有效的过滤背景数据，从而成千倍的节省数据量，降低系统成本，并使实时的手势识别变得很容易。

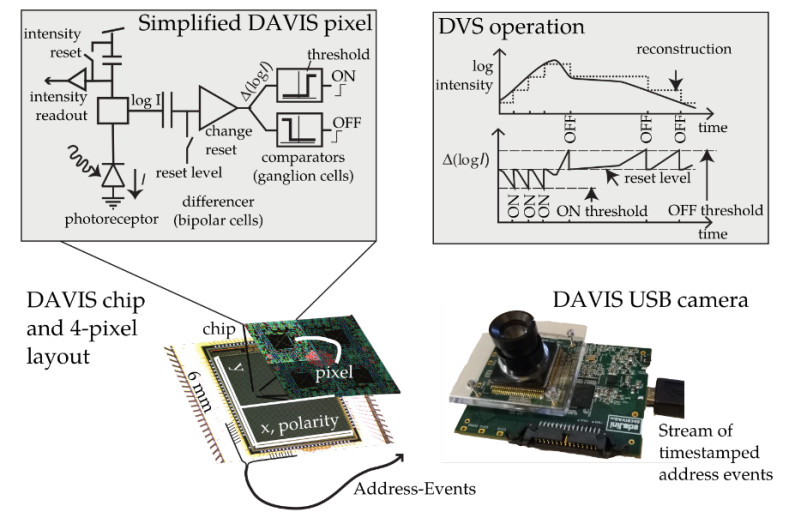
# 2．国内外研究现状

## 2.1 动态视觉传感器研究现状

近年来， 计算机视觉领域出现了一种新的相机模型——事件相机，也称为动态视觉传感器(Dynamic Vision Sensor, DVS)。 DVS 是受生物启发的传感器，其工作原理与传统相机截然不同。 不是以固定的速率捕获图像，而是异步测量每个像素的亮度变化， 输出编码了事件、 位置和亮度变化极性的事件流。 DVS128[1]是第一款商用的 DVS 相机。

传统的相机传感器是基于帧进行成像， 在固定的间隔(曝光)时间内，像素阵列检测到的光电流被积分在电容器中，在每一帧时间内，每个像素的积累的电压电平以顺序方式传送出芯片。因此使用传统相机传感器时，每一帧都会将所有的信息传输出去，不管单个像素上是有变化进而判断是否需要进行信息传输。此外，由于光电流是在固定时间段(通常在20-30ms)完成积分，所带来的延迟对于高速场景会造成模糊。

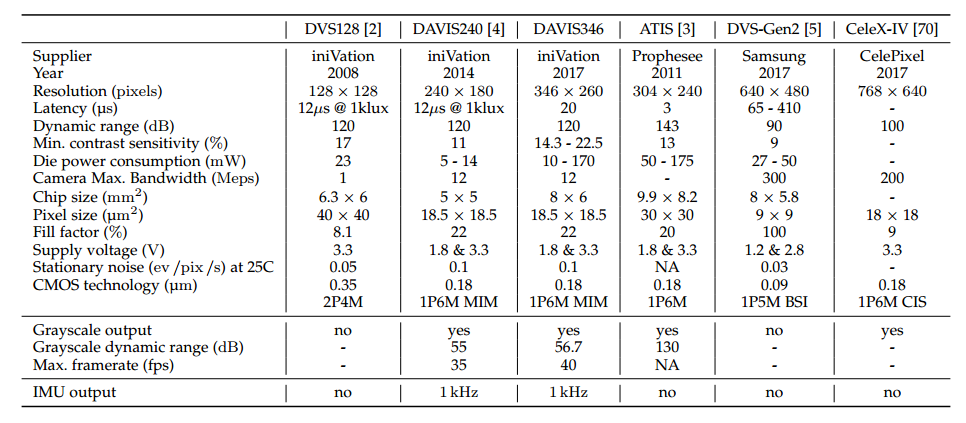
标准相机获得完整的图像速度由外部时钟指定(如30 fps),事件摄像机,如动态视觉传感器(DVS),异步响应场景中亮度变化和独立的每个像素[2]。图一是DAVIS相机的简易示意图，左上图是每一个DVS 传感器像素的简化示例，每个像素记录着对数光强，当超过了一定的阈值之后就发射一个事件， 包含了坐标、事件以及 1 位的极性，“ON”表示亮度增加，“OFF”表示亮度降低。 右上图展示，事件相机的输出变量是包含事件和峰值的数字序列，每个事件表示在特定时间内一个像素除的强度变化。目前使用最广泛的事件相机像素较大:DVS128、ATIS、DAVIS240、DAVIS346、ATIS等。表1显示了当前流行的事件相机的具体参数。



DAVIS 相机简易示意图[2]

DVS 相机相比于传统相机， 具有杰出的属性：非常高的动态范围(140dB VS. 60dB), 非常高的时间分辨率(us 数量级)，更低能耗， 没有动态模糊。 因此，在高速、高动态范围等传统相机面临挑战的场景中， 动态视觉传感器在计算机视觉方面更具很大的潜力。然而，需要新的方法来处理这些传感器不同于普通相机的输出图像，以释放它们的潜力。

事件相机是一种革命性的传感器，与传统的基于框架的相机相比，它具有许多优点，如低延迟、低功耗、高速和高动态范围。因此，它们在计算机视觉应用方面有很大的潜力，基于事件的视觉是一种新兴的视觉技术，可以用于目前传统相机无法达到的具有挑战性的场景。



商用 DVS 相机参数[2]

## 2.2 手势识别研究现状

手势识别是人机交互研究领域的一个重要分支，是机器视觉领域的一个重要研究内容，其发展状况与人机交互技术紧密相连。成熟的手势识别产品已在世界范围内广泛使用，典型的手势识别类电子设备有：2003年，Sony公司推出一款名为EyeToy的手势识别设备，这种设备能够将玩家的动作传输到游戏画面，实现与玩家的互动；2010年，微软公司推出的Kinect体感设备在手势跟踪与识别方面有着出色的表现，它能实时识别用户手势，结合Xbox，是用户完成对游戏的控制指令；2012年，三星推出的智能电视ES8000结合用户的手势，可以对电视进行换台、搜台以及音量调节等操作；2016年，宝马7系引入了手势识别功能，驾驶员可以调高或调低音量，接听或拒绝电话等。

近几年的手势识别的研究，大多围绕着传统相机或者深度相机进行。Christian Zimmermann和Thomas Brox[1]提出了一个从常规RGB图像中估计三维手势的方法，建立了一个深度卷积网络，在此之前学习网络隐式3D关节。 连同图像中检测到的关键点一起，该网络可以很好地估计3D姿态。Christian Zimmermann和Thomas Brox引入了基于合成手模型的大规模3D手姿势数据集，用于训练所涉及的网络。 在包括手语识别在内的各种测试集上进行的实验证明了在单色图像上进行3D手势估计的可行性。英伟达公司的Pavlo Molchanov等[2]提出了基于循环三维卷积神经网络的动态手势在线检测与分类，实现了实时手势识别，避免了在执行手势及其分类之间出现明显的滞后。手势视频以短片段的形式呈现给3D-CNN网络，用于提取局部时空特征。 这些提取出来的时空特征输入到循环网络中，该网络汇总了多个片段之间的过渡，最后通过softmax函数进行分类。西安电子科技大学的张亮等[3]通过在卷积网络LSTM基础上加上注意力机制，提出了“RES3D+ConvLSTM+Mobilenet“网络结构，并对ConvLSTM提出的变体在大规模孤立手势数据集Jester和IsoGD上进行了评估，取得极佳效果。

基于动态视觉传感器的手势识别研究这几年也越来越多。IBM研究所的Arnon Amir等[4] 提出了首个端到端基于事件的硬件上实现的手势识别系统，该系统使用TrueNorth神经突触处理器以低功耗实时地从动态视觉传感器（DVS）流式传输的事件中实时识别手势。Arnon Amir等首次使用TrueNorth处理实时DVS事件流，TrueNorth是具有100万个尖峰神经元的基于事件的本地处理器。 TrueNorth芯片在这里配置为卷积神经网络（CNN），以105 ms的延迟识别手势的开始，而消耗的功率却不到200mW。 CNN在一个新收集的DVS数据集（DvsGesture）上达到了96.5％的样本外准确度，该数据集包含3种光照条件下来自29个对象的11种手势类别。由于DVS具有功耗低、时间分辨率高、动态范围大、存储需求少等优点，Stefanie Anna Baby等[5]探讨了使用DVS进行人类活动识别（HAR）的可行性。他们提出使用DVS视频的各种切片作为HAR的特征映射，并将它们表示为运动图。他们发现将运动地图与运动边界直方图（MBH）融合，在基准DVS数据集以及我们收集的真实DVS手势数据集上具有良好的性能。

## 2.3 发展趋势

机器视觉已经越来越多地代替人类进行视觉信息的获取，帮助处理复杂的现实生活环境。靠着算法的提升，机器视觉的处理速度、跟踪物体准确能力已经有了很大的改善，但是在处理复杂的现实环境时，还是会有误差大、计算量大、功耗大的问题。模仿人类脑神经元的卷积神经网络打造出了强大的人工智能，模仿人类视网膜结构也应当能给机器配备强大的视觉能力。

在动态视觉传感器中，对于单个像素点，只有接受光强度发生改变时，才会有事件(脉冲)信号输出。DVS可以拍摄高速运动的物体，即便物体运动非常快，也能拍摄清楚并且以任何速度播放。而且它比高帧率摄像机更能胜任这项任务，不会如同基于帧率的摄像机那样产生大量的冗余数据，对运算资源、能耗造成浪费。

与基于帧的摄像机相比，基于事件的动态视觉传感器通过在像素检测到亮度变化时发送异步事件来模仿生物视网膜，从而消除了多余的数据传输。用户做出的手势事件经过动态视觉传感器处理后传递到卷积神经网络进行手势识别。

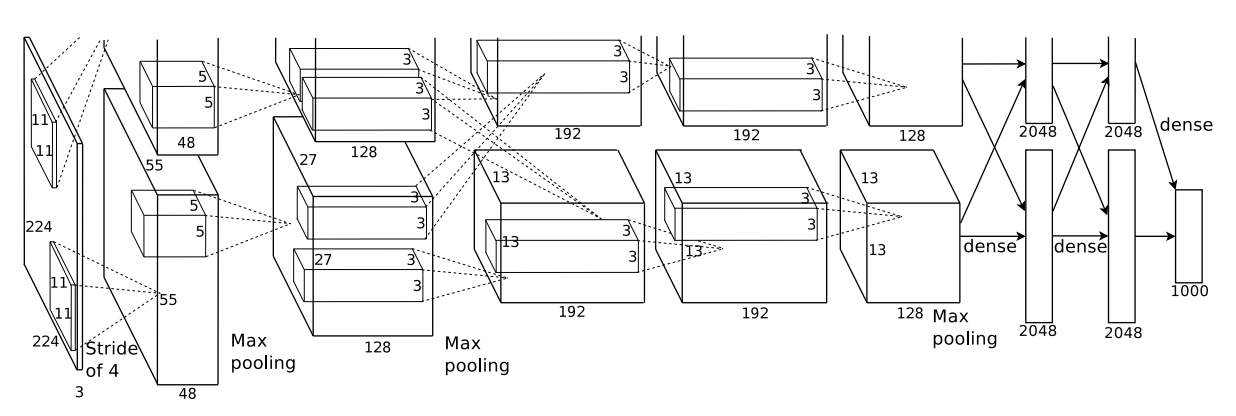
# 3．相关技术原理和技术路线

首先利用动态视觉传感器收集DVS数据集，完成数据的成帧处理。对于静态的手势识别，处理得到的数据集是各个手势（例如剪刀、石头、布）的单个图片，将数据传入AlexNet网络[7]进行训练；对于动态的手势识别，处理得到的数据集是对应动作（例如左挥手、右挥手、抖手等）连续的图片或视频片段，将数据传入由3DCNN和LSTM所搭建的网络进行训练。

## 3.1．技术原理

### 3.1.1 静态手势识别神经网络训练算法（AlexNet）

Alex在2012年提出的AlexNet网络[]结构模型引爆了神经网络的应用热潮，并赢得了2012届图像识别大赛的冠军，使得CNN成为在图像分类上的核心算法模型。在此之前，神经元的激活函数一般选择sigmoid或者tanh，然而在AlexNet网络中选择的是ReLU(修正线性单元),用这个激活函数其训练速度要比传统的神经网络快几倍。



AlexNet 该模型一共分为八层，5个卷积层,，以及3个全连接层,在每一个卷积层中包含了激励函数RELU以及局部响应归一化（LRN）处理，然后在经过降采样（pool处理）。从**图上**可以明显看到网络结构分为上下两侧，这是因为网络分布在两个GPU上，这是因为NVIDIA GTX 580 GPU只用3GB内存，装不下那么大的网络。为了减少GPU之间的通信，第2,4,5个卷积层只连接同一个GPU上的上一层的kernel map(指和卷积核相乘后得到的矩阵)。第三个卷积层连接第二层的所有的kernel map。全连接的神经元连接到上一层的所有神经元。第1,2,5卷积层卷积之后进行max pooling操作。ReLU激活函数应用于所有的卷积层和全连接层。

### 3.1.2 动态手势识别神经网络训练算法

3D-CNN[8]能够直接从原始输入中提取特征，通过执行3D卷积在视频中从时间和空间维度提取特征，将高级功能模型规范化，并结合各种不同模型的输出，进一步提高3D CNN的性能。3D-CNN在视频分类，动作识别等领域发挥着巨大的优势。3D卷积是通过堆叠多个连续的帧组成一个立方体，然后在立方体中运用3D卷积核。通过这种结构，卷积层中的特征图都会与上一层中的多个相邻帧相连，从而捕获运动信息。

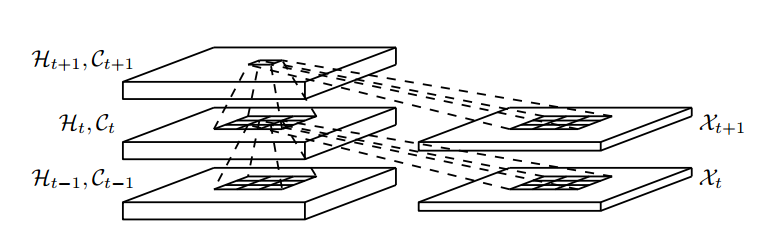
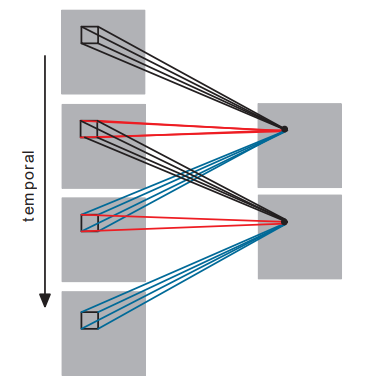


Figure 3 3D-CNN

长短期记忆（Long short-term memory, LSTM）[]是一种特殊的RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说，就是相比普通的RNN，LSTM能够在更长的序列中有更好的表现。ConvLSTM[]核心本质还是和LSTM一样，将上一层的输出作下一层的输入。不同的地方在于加上卷积操作之后，为不仅能够得到时序关系，还能够像卷积层一样提取特征，提取空间特征。这样就能够得到时空特征。并且将状态与状态之间的切换也换成了卷积计算。

在处理时空数据方面，LSTM的主要缺点是，它在输入到状态和状态到状态的转换中没有对空间信息进行编码。为了克服这个问题，我们设计的一个显著特点是所有的输入X1;::;Xt，单元输出C1;::;Ct，隐藏状态H1;::;Ht，和门 是三维张量。为了更好地了解输入和状态，我们可以把它们想象成空间网格上的向量。如图【】所示ConvLSTM通过其本地邻居的输入和过去状态来确定网格中某个单元的未来状态。在状态到状态和输入到状态的转换中使用卷积算子可以很容易地实现这一点。ConvLSTM的关键方程如下【】所示，其中为【】卷积算子，【】为Hadamard乘法。

这种结构也类似于LSTM预测模型，只是我们的输入和输出都是保存所有空间信息的三维张量。由于该网络具有多层叠加的ConvLSTM层，因此具有较强的表示能力，适用于处理动态的手势动作。

## 3.2 技术路线

基于动态视觉传感器（DVS相机）的手势识别系统的体系架构设计流程如图。首先解决DVS数据处理，包括成帧、去噪等步骤，获得网络训练所需要的数据集。然后分别通过AlexNet网络和“3D-CNN+LSTM”网络进行手势识别。

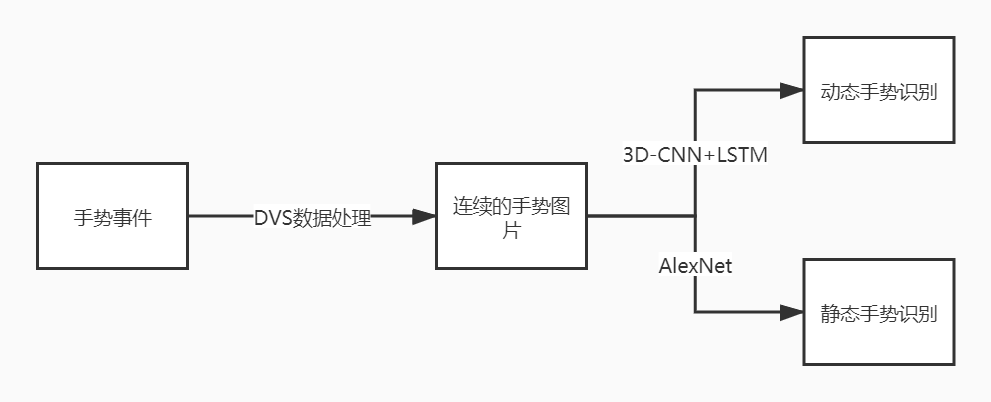
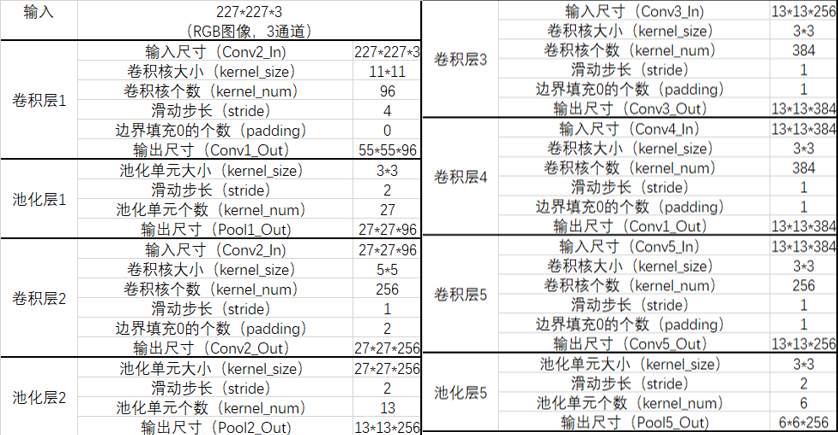


Figure 4 技术路线图

3.2.1 DVS数据处理

3.2.2 静态手势识别

我们的目标是获取到静态的手势图片，实现分类。为了实现我们的目标，我们采用现有的AlexNet网络。使用DVS数据处理得到的DVS数据集进行训练。





通过AlexNet网络对物种静态手势（剪刀、石头、布、OK和点赞）进行训练，最后输出的分类结果只有五种，需要对AlexNet网络的全连接层进行调整，将全连接层的4096个过滤器减少为1024个过滤器，最后输出5种分类结果。上图显示了我们修改后的AlexNet网络结构。

3.2.3动态手势识别

# 4．研究难点及目前存在的问题

## 4.1 研究难点

1. DVS数据的处理。DVS传送的事件需要经过去噪、渲染、成帧等处理才能转换成神经网络训练所用的数据类型。

2. 实时手势识别网络的搭建。如何将3D-CNN与LSTM相结合，网络参数的确定。

3. 降低在线的动态手势识别的延迟。对于动态手势识别，需要尽可能减小做出手势与调用模型识别之间的延迟，以实现实时效果。

4. 将实时手势识别和基于事件的手势识别结合。虽然实时手势识别和基于事件的手势识别的研究很多，但是二者结合的相关研究很少。

## 4.2 存在的问题

在实际的研究和设计过程中，存在如下几个问题：

1. DVS数据的成帧和去噪处理；

2. 动态手势识别网络的优化；

3. 调用模型识别手势的延迟。

# 5．研究预期成果

1. 基于动态视觉传感器（DVS相机）的实时数据显示。通过动态视觉传感器实时捕捉突然出现的人物目标，实现实时数据显示，同时利用动态视觉传感器制作手势图数据集，并开源数据集；

2. 搭建完整的基于动态视觉传感器的（DVS相机）的手势识别系统，实现静态的手势分类和动态的手势识别；

3. 在完成毕业论文的同时，发表一篇期刊论文。

# 6．进度安排

1. 2020.01.10-2020.03.04 查阅文献，提交开题报告，确定文章的大体结构；

2. 2020.03.04-2020.03.30 完成论文完整提纲，进一步收集相关资料及文献，同时完成DVS数据处理和手势识别相关工作；

3. 2020.04.01-2020.05.10 按照毕业设计任务书要求开展研究工作，完成论文的初稿，并与导师商讨不合理之处；

4. 2020.05.10-2020.05.30 修改论文并定稿，提交导师检查；

5. 2020.06.01-2020.06.15 排版、打印装订，上交论文，准备答辩。

参考文献

1. Gallego G, Delbruck T, Orchard G, et al. Event-based vision: A survey[J]. arXiv preprint arXiv:1904.08405, 2019.
2. P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128× 128 120 dB 15 µs latency asynchronous temporal contrast vision sensor,” IEEE J. Solid-State Circuits, vol. 43, no. 2, pp. 566– 576, 2008.

[1] ChristianZimmermann, Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Image[J]. 2017:4913-4921.

[2] Molchanov P , Yang X , Gupta S , et al. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks[C] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[3] Zhu Guangming,Zhang Liang,Yang Lu,Mei Lin,Shah Syed Afaq Ali,Bennamoun Mohammed,Shen Peiyi. Redundancy and Attention in Convolutional LSTM for Gesture Recognition.[J]. IEEE transactions on neural networks and learning systems,2019.

[4] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah. Attention in Convolutional LSTM for Gesture Recognition[C]. NIPS,2018.

[5] Amir A , Taba B , Berg D , et al. A Low Power, Fully Event-Based Gesture Recognition System[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[6] Stefanie Anna Baby, Bimal Vinod, Chaitanya Chinni, Kaushik Mitra. (2018). Dynamic Vision Sensors for Human Activity Recognition. 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)

[7] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]. NIPS, Curran Associates Inc, 2012.

[8] Ji, Shuiwang et al. “3D Convolutional Neural Networks for Human Action Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2010): 221-231.

[]Hochreiter, Sepp, Schmidhuber, Jürgen. Long Short-Term Memory[J]. Neural Computation, 9(8):1735-1780.

[]Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. Computer Science,2015.